What are the Needs when working with Semantic Resources?

A collection of use cases in agriculture and nutrition

Last update: June 28, 2018

The use cases presented in this document where collected by the Agrisemantics Working Group of the Research Data Alliance (https://www.rd-alliance.org/groups/agrisemantics-wg.html) during spring 2018.

A report on the methodology adopted and the analysis of the use cases is available on the group website.

We thank all the persons who shared their use cases and requirements with the group.

Table of contents

UC01: Search data set by the semantic resource it uses by Valeria Pesce (GFAR)

Problem statement

We run a catalogue of "standards", data and metadata (the Vest Map of agri-food data standards), and a catalog of datasets (the CIARD RING). In the catalog, we want to support a search by the semantic resources used in a dataset, but the metadata describing a dataset (harvested from other existing catalogs) seldom includes this and it is hard for us to extract this piece of information directly from the dataset. Ideally, this piece of info should be included in the metadata scheme used to describe the dataset, and automatically generated when creating the metadata. People should also be encouraged to use it. The problem obviously also appears when harvesting metadata from other sources.

Problem 1. DCAT allows for the use of dct:conformsTo, but this is not very well known. Major repository services such as Dataverse and CKAN do not include such type of metadata. We would like a property like dct:conformsTo to be always included in datasets metadata.
Problem 2. Besides, the value of the dct:conformsTo property is a Resource, so there would also be the need for an agreed RDF authority list of data standards (which in our case would be the Map of agri-food data standards).

Ontologies and vocabulary requirements

A. Can you say something about the semantic resources you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

Any, although the focus here would be on schemas and ontologies that describe datasets

B. How are they used? What for?

For describing datasets

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

It depends. DCAT is publicly accessible; other schemas used by other tools have never been published.

D. Do they come with a clear license? Which license?

It depends

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

It depends

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

The main difficulty is that relevant vocabularies and relevant tools are not under our control. DCAT would be sufficient if the dct:conformsTo property were used. The real problem is data repository

tools either adopt DCAT partially or use their own internal XML or Json schemas which do not include such a property.

The only possible solution seems to be advocating the use of full DCAT to the producers of data repository tools. (This would actually solve many other dataset metadata interoperability issues beyond this specific use case.)

Tell us a bit more about…

Data requirements

A. What type of data, format, storage, size, workflow, etc apply to this use case?

Type of data: dataset metadata

Format: any RDF format (but even non-RDF XML or Json)

Workflow: dataset management

Manpower

A. What is the profile of persons typically encountering the problem you described.

The person who manages a dataset catalogue and wants to implement the search by data standard.

Organization and role

What is your organization and your role in it?

GFAR, Information Systems Manager and Project Manager

UC02: *Metadata Exploitation* by Armando Stellato (Tor Vergata University)

Problem statement

Currently many datasets are available and linked on the LOD, yet their retrieval is limited to a few services (such as LOV or Datahub.io) and their nature (ontology, thesaurus, datacube?, which lexicalization model they adopt?) is not exploited by data retrieval, alignment, browsing systems in order to facilitate their access.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

All of the above.

B. How are they used? What for?

When it is about thesauri, logical axioms are sometimes available in their defining vocabularies order to better describe the resource, so constraining it (useful for editing tools) rather than for reasoning.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Well, I'm a Knowledge Engineer, and as a consultant, I tend to recommend publication on the LOD. Some of them though, sometimes are not, or are only partially available.

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

Usually there is no jack-of-all-trades. Usually adopted tools involve:

- Editing
- Publication as LD
- Browsing

The second and third might come together as a one-solution or be split in different, more specific solutions.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Adoption of Metadata. There are standards out there, yet there is not much exploitation of them around in existing systems and this is need for improving and automatic resource discoverability and reachability (and for improving the quality of the interaction with the resources by knowing their characteristics)

For what concerns my role in the matter, I'm leading the development of VB3 and we have already provided metadata production functionalities for VB3, whereas on the roadmap for this year, we will check how to improve the metadata exploitation in order to reach resources on the web, to align them etc.. and to discover the best ways to do that.

Besides my specific contribution, I would like to see more metadata exploitation in the whole scenario of linked data both at the production and consumption level.

Tell us a bit more about…

Organization and role

What is your organization and your role in it?

University of Rome, Tor Vergata

Researcher

UC03: VocInra: turning an institutional keyword list into a linked open thesaurus by Sophie Aubin (Inra)

Problem statement

The INRA Department of Scientific Information (DIST) works transversely to support INRA scientific strategy and provide innovative services within the area of data, information and

knowledge management. DIST develops tools and services to support the research activities, and promote open access to scientific and technical information.

DIST maintains the institutional reference vocabulary VoCInra that is used internally by STI professionals and researchers to index scientific textual content.

VocInra has grown over years in a rather uncontrolled way and thus needs some cleaning and rearrangement. The structure is unbalanced across sections and the whole vocabulary do not answer the user needs any longer neither the current both technological and quality requirements.

Our aim is to improve the quality of VocInra and to conform it to the open science principles adopted by Inra, namely make it FAIR, and linked to external open semantic resources, GACS being the priority.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

VocInra contains approx. 93,000 terms, either descriptors in French, synonyms or equivalents (mainly in English). They are organised into 8 thematic entries and present approx. 10,000 hierarchical relationships. The existence of a similar concept in external vocabularies like Agrovoc is mentioned but no explicit reference to the concept is provided which would allow to create a link to it.

We would like to interlink VocInra with GACS.

B. How are they used? What for?

VocInra is used internally by STI professionals and researchers to index scientific publications in the institutional archives (ProdInra) and describe the activities of labs and people in the institutional directory (http://annuaire.inra.fr).

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

VocInra is local and has never been published (but we intend to).

D. Do they come with a clear license? Which license?

No

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

VocInra is maintained in an internal tool based on a relational database. The data format is ad hoc. The interface allows to query by anyone, add, edit and remove entries (by the manager only). It is connected through web services to two applications that use VocInra, namely ProdInra and Annuaire. Additions to VocInra can be asked by users in the ProdInra interface directly. The VocInra manager accepts or rejects the demands. The tools does not allow for browsing the vocabulary, extracting sub-parts of it or performing mass edits. It is not possible to easily move an entry inside the vocabulary. In a word, edit functionalities are really limited.

We developed an XSLT script to transform it from internal format to RDF/SKOS-XL.

We experimented Onagui (https://github.com/lmazuel/onagui) to create some alignments from VocInra to Agrovoc in the domain of agroecology.
All tools are standalone which makes the whole process complicated and barely sustainable.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

It is difficult to find the tool that will allow us reshaping VocInra and perform the quality improvement work. We tried VocBench2 but faced too many difficulties due to the specific structure of our vocabulary (very flat). DIST does not control the VocBench installation which is managed by another Inra entity. There are uncertainties about installing VocBench3.

Having a service providing access (and support) to the last version of VocBench with the possibility to create private projects would be valuable.

Such an editor should include quality assessment functionalities to check hierarchy circles, redundant entries, equivalents missing, etc.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Bibliographic data mainly conforming to OAI-PMH formats.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

People maintaining the resource are knowledge engineers and librarians. The users are librarians or researchers who interact “manually” with the vocabulary through the two applications.

Organization and role

What is your organization and your role in it?

Inra

I am a knowledge manager in charge of developing services and vocabularies for information management or text mining purposes dedicated to researchers.

UC04: *Linking Wheat Data With Literature* by Robert Bossy and Thomas Letellier (Inra)

Problem statement

The objective of the use case is to improve the FAIRness (Findable,Accessible, Interoperable, Reusable) of international datasets. It allows one to find wheat datasets, including, genomic, genetic, phenotyping, and bibliography in an international network. Data discovery is achieved by identifying shared resources between bibliography annotated in the Openminted platform and datasets and in particular those hosted in GnpIS, the INRA-URGI integrative data repository. Those shared resources include traits, markers, genes, and plant variety description. Two critical elements of this data

discovery system are GnpIS and an international distributed indexing mechanism, the WheatIS search. Both are currently developed and maintained by the INRA-URGI platform which belongs to Elixir-fr (or IFB, Institut Français de Bioinformatique) French node of Elixir, the European eInfrastructure for Bioinformatics. **The text mining workflow processes scientific literature and textual fields of the GnpIS database in order to identify and annotate taxa, markers and phenotypes that are described and accessible in both applications. The parent applications and the text mining workflow use two different ontologies. To allow interoperability between datasets, mapping between concepts of both ontologies are created.**

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

The **WIPO Wheat Phenotype Ontology** is the ontology used at Inra URGI to reference the phenotypic information described and displayed in the GnpIS and WheatIS search applications. It includes definitions for approx. 260 wheat phenotypic traits in terms of variable, growth stage, measurement method, and scale. It is organized following the Cropontology framework in trait types, e.g. *physiological trait*. It contains synonyms for traits. It is built and versioned using the Cropontology spreesheet format which can be converted into BreedingAPI web services format and into OWL through the cropontology toolset.

The **Wheat Trait Ontology**, developed and maintained at Inra Maiage contains approx. 466 wheat phenotypes and environmental factors that impact wheat varieties. Built for text mining purposes, it lists and organizes concept labels that are close to the terms that are usually found in literature, including many synonyms. The concepts are organised hierarchically in a obo file. The use case uses only the part with phenotypes.

The **NCBI taxonomy** is used to annotate taxon names. The taxonomy contains the taxa from the phyla to sub-specific ranks, the tree structure, and the most common synonyms (~1.4M taxa, 2M names).

A **lexicon for markers** is provided by URGI as a tabulated list. Each entry includes a unique identifier, a preferred name and synonyms.

B. How are they used? What for?

The **WIPO** is used by applications to identify and describe the data related to traits referred in genomic, genetic, and phenotyping datasets published at URGI. It ensures long term accessibility and interoperability. It is also used to increase findability.

The **Wheat Trait Ontology** is used in the text mining pipeline for the recognition and the categorization of the wheat phenotype mentions in the text.

The NCBI taxonomy and lexicon for markers are used similarly for taxa and markers respectively.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

WIPO is publically available and can be downloaded at this address: https://urgi-git.versailles.inra.fr/urgi-is/ontologies/tree/develop/Wheat
It is used by national and international consortia.

WheatPhenotype ontology is publically available in AgroPortal: http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE.

The **NCBI taxonomy** is publically available at ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz

The **lexicon for markers** is not publically available.

D. Do they come with a clear license? Which license?

WIPO: CC-BY-SA license v4.0
Wheat Trait Ontology: CC-BY-SA license v3.0
The **NCBI taxonomy** is under UMLS Metathesaurus License

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

- Editing, maintaining the ontologies

At URGI, Excel is used to edit **WIPO** following the Crop Ontology template and framework. Its version are followed in a Git versioning system (Gitlab: https://urgi-git.versailles.inra.fr/urgi-is/ontologies).
At Maiage, the **WheatPhenotype ontology** is edited using Obo-Edit2 as it is the most convenient for domain experts (much simpler interface than Protégé).

- Using the ontologies

At URGI, **WIPO** is used by a widget in GnpIS to make queries by variables (e.g. https://urgi.versailles.inra.fr/ephesis/ephesis/viewer.do).
The **WIPO** is also used in the WheatIS Data discovery search for the phenotype data type**.** The records are enriched with concepts from **Wheat Trait Ontology** when a link exists between a **WIPO** concept and a **Wheat Trait ontology** concept.

At MaIAGE, a complex text mining pipeline, Alvis, is integrated into the OpenMinTeD infrastructure. It includes a lexicon projection algorithm that takes text, and the **NCBI taxonomy** or the **marker lexicon** as entries. The output is the text annotated with identifiers for taxa and markers. The annotation of phenotypes is performed in two steps: 1) a term extractor (YaTeA) extract terms according to linguistic and statistical criteria, then 2) a normalizer (ToMap) maps the extracted terms with the concepts of the **Wheat Trait Ontology**. The resulting annotated text is enriched with concepts from **WIPO** for which links exist with those in the **Wheat Trait Ontology.**

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

- Editing

Using Obo-Edit2 represents risks as it is not maintained any longer. Protégé cannot be a good alternative as it is too complicated for non ontologists. Ontology developers need easy interactions with domain experts (direct editing, comments, validation, etc.)

The Cropontology/URGI editing workflow is efficient and biologist friendly. It could be improved by adding OWL transformation following the work initiated in the RDA RDFENO project and following the outputs presented in http://ist.blogs.inra.fr/wdi/phenotypes-as-rdf/

- Using

In text mining environments, there is no standard to represent the annotation of text elements with ontology concepts. Some proposals have been advanced like XMI (used by UIMA) or Open Annotation, but none has become a de facto standard, due to the fragmentation of the text-mining community and the tension between specifications (performance vs. shareability). As a result each text-mining project shares annotations in an ad hoc format.

The ontologies used are well organised, thus allowing rich annotation with measurement methods and more general mapping and searching through the trait. For search, there is currently a development to inject WIPO and other ontologies in the distributed search engine. It is funded through Elixir-FR/IFB, but more fundamental research approaches development on the best way to do fuzzy search on trait names and descriptions, or on using the entity (e.g. “leaf”) and its attributes (e.g. “area” to get the “leaf index index”) might help the progress on this distributed search functionality.

- Aligning

The alignment between concepts from **WIPO** and **Wheat Trait Ontology** is done manually in an Excel sheet involving several exchanges between experts. We clearly lack a dedicated tool to animate and document this process. The issue is not the number of concepts and alignments here but quality checking and collaborative features instead.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

On the one hand, wheat phenotypic data which some are in open access (Small grain cereals network) and the others will be available at the end of the projects (BreedWheat, Whealbi). Those data are MIAPPE (www.MIAPPE.org) compliant and stored in GnpIS. They can be retrieved in several MIAPPE implementations : ISA Tab archive, CSV files, BrAPI web services and RDF.

On the other hand, bibliographic (metadata) and textual (full text) data. This use case aims at processing tens of thousands of bibliographic records and as many corresponding full texts as available online.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

At URGI, the persons maintaining and using the WIPO are data managers, application developers. The work closely with domain experts to build and maintain them.

At MaIAGE, the persons maintaining **Wheat Trait Ontology** and using other semantic resources in the text mining pipeline are computational linguists and machine learning scientific experts.

Organization and role

What is your organization and your role in it?

Robert Bossy is in charge of the Alvis plateform and of its integration into the OpenMinTeD plateform.

Thomas Letellier is data manager at Inra URGI, in charge of wheat data integration.

UC05: Data integration for sensory and environmental quality in food by Liliana Ibanescu (AgroParistech and Inra)

Problem statement

The research from the "Delicious" project described in this use case aims at producing well-balanced products, here dairy gels, in terms of nutritional requirements (e.g. less fat, salt, and sugar) while using eco-friendly transformation processes. This research will allow to address two societal challenges which are population ageing, and overweight and obesity. This requires an advanced analysis of the production process according different criteria, and relies on the combination of heterogeneous data and knowledge (on food composition digestion, digestion, sensorial perception, food processing, etc). The integration of the data with a great variety in types, vocabularies, and formalisms, is realized with the help of an ontology: PO² (Process and Observation Ontology).

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

PO² (Process and Observation Ontology), developed and maintained by us, describes a production process composed of steps and having different possible itineraries. A production process transforms a mixture using the combination of a material and a method. An observation observes different attributes for a step or a mixture during a given step. Available data concerns the production of cheese including production process description, composition and structure of the studied cheese during the different steps, and sensory perception parameters.
The core ontology has 90 classes and 122 properties. They are described with English labels, definitions, and notes.

B. How are they used? What for?

PO² is used in combination with an EXCEL template which allows experts to record data. This data is transformed then stored in a triple store which can be queried through a sparql endpoint.
In the future, the ontology may be used to better control the allowed values in the EXCEL template.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

The core version of PO² (version 1.5) is publically available on AgroPortal: http://agroportal.lirmm.fr/ontologies/PO2/?p=summary .

D. Do they come with a clear license? Which license?

Under CC-BY 4.0

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

CMap tool was used for the conceptualisation of the ontology. It allowed collaborative work and discussion with the experts.

The core component of PO² was developed using Protégé.

An in-house tool makes the conversion from EXCEL sheets into RDF data to be stored in a triple store. Data are then searched and extracted using sparql queries.

Alignments with Agrovoc, NALT, and GACS are computed using an in-house tool, PO2VocabManager developped in Java.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Further specialisation of the core components to specific domains cannot be made in Protégé, too complex and too permissive. Therefore our tool PO2VocabManager was designed as a specific interface adapted for a non ontologists, i.e. domain experts who can easily use it with some limited editing rights in order to reduce damage risks.

External endpoints could be used to get additional data and knowledge but they are sometimes not accessible, or difficult to identify.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Input domain data are heterogeneous and stored in journal papers, PhD thesis, EXCEL files. The structure imposed by PO2 ontology allows to build a Triple store.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Software engineers,, computer scientists expert in knowledge representation.

Domain experts specialists in food.

Organization and role

What is your organization and your role in it?

Computer scientist expert in knowledge representation at AgroParistech and Inra

UC06: Improve management, application, validation of terminologies at Embrapa, and training on using them by Ivo Pierozzi Jr (Embrapa)

Problem statement

Embrapa has an extensive knowledge base of publications and data used internally (researches, technicians, librarians and managers) and also externally by users and customers, though services and technologies, as well as the academic and scientific community;

For its internal use, Embrapa developed a controlled vocabulary, AgroTermos. This needs to be consolidated in its terminological component (Brasilian Portuguese), and corporate and operational processes for its construction and management should be defined and implemented. The goal is to develop AgroTermos so that it provides the conceptual model of Brazilian agricultural knowledge representation. We also need to test applications consuming AgroTermos in corporate information systems, evaluate their performance in contributing to improve semantic interoperability and in supporting indexing and retrieval of digital information.

The expected impacts of AgroTermons are the following:

- better conditions for mapping, organization, recombination and dissemination of Brazilian agricultural data and information;
- better visibility of Brazilian agricultural scientific production (better insertion of the Portuguese language in the international scenario of agricultural knowledge);
- better interoperability between internal and external information systems used or developed by Embrapa;
- better conditions to exploit technological and semantic resources for the representation of agricultural knowledge and its valorization in the development of artificial intelligence and its insertion in the digital agriculture

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

AgroTermos was initially constructed starting from the commonalities of the Portuguese terminologies already existing in the agricultural thesauri Agrovoc, CABt and Thesagro. It is intended to reuse the methodologies, ICT and good practices used for the construction of the GACS, with which AgroTermos should interoperate via webservices.

They include all of this: *hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.*

B. How are they used? What for?

AgroTermos intend to be a “conceptual schema” including some additional terminological informations like definitions, linguistic and grammar properties or context for each term, aiming at the construction of digital glossaries and dictionaries for agricultural knowledge domains.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

It is intended that the management process of AgroTermos be local, but that its products and services would be accessed publicly. URL and URI should be shared.

D. Do they come with a clear license? Which license?

This depends on a coaching/benchmarking consultation of skills already existing in systems of the same type when we can learn the best practices of licensing. But where it is possible to use, depending on the legal possibilities at Embrapa, it is intended to use free licenses (Creative commons?)

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

General tools:

Cmap tools; yEd; Gephi; NodeXL; Qiqqa; Antconc; VOSViewer; Cortext tools; Pajek; Qiqqa

In-house tools:

e-Termos; Etecam; Embrapa namespace; conceptual structures visualization tools; textual similarity algoritms

Some theoretical references:

Knowledge mapping; semantic mapping; corpora linguistics; KOS progressive model…

Obs.: no integration of them!!!

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Difficulties:

- Brazilian Portuguese language and its insertion in the global context of agricultural knowledge;
- Great volume of data to processing;
- Ontology merging methodologies;
- Low index of integration of methods and technologies in the complex workflow;
- Medium to long-term project execution, but it can be run in "module" or "phases";

Bottlenecks:

- Financing;
- Institutional/political articulations aiming partnerships, intellectual property (= annoyant bureaucracy)

Solutions:

- AgroTermos semantic model adjustment;
- Procedural model for new terms to be included in AgroTermos;
- Interlinking AgroTermos with GACS
- (…)

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Perhaps the answers to this question is exactly what we are trying to answer with the proposal of this use case (!!!)… but we are working with structuring textual data as "terminologies" and from there

constructing KOS where semantic resources can be incorporated in models structurally and functionally more adequate to the representation of agricultural knowledge and its applications

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Terminologists (linguists), information science professionals, domain experts, computer professionals; software developers, librarians; journalists; corporate managers Domain experts specialists of food bring their knowledge.

Organization and role

What is your organization and your role in it?

Embrapa Informática Agropecuária

Knowledge Mapping and Organization & Agricultural Terminology

Embrapa Agricultural Informatics

UC07: Link and search ontologies and vocabularies to achieve semantically-driven assessment of economic returns from biodiversity protection by Ferdinando Villa (Ikerbasque Center for Climate Change)

Problem statement

In the context of an web-based platform (ARIES) oriented to the study of ecosystem services, ontologies and vocabularies are widely used as a way to access and reason on third-party datasets. Below is a narrative example of a query that ARIES can support:

“Estimate the differential in agricultural yield attributable to increase or decrease in pollination caused by the protection of a given proportion of natural habitat in any region where agriculture is practiced and pristine natural habitats exist.”

All data accessed to answer that query come from semantically annotated, distributed spatial databases. The user is not expected to define any parameter except the geographic and temporal contexts for the query.

The process implemented to answer the query can be summarized as follows:

1. Find and retrieve spatial data for agricultural crops and biodiversity based on core ontology.
2. From data or additional queries, determine the specific identities of each crop, pollinator population and keystone species population involved the context, based on linked authorities/vocabularies.
3. Find and retrieve tables of pollinator ranges and behavior based on identities and context.
4. For each scenario (at minimum baseline and increased protection):
    a. Compute and create pollination supply areas based on natural features;
    b. Link to existing crop areas based on dependence of crop from pollinators, pollinator range and regional biophysical characteristics (elevation, humidity etc);

    c. Assess yield proportion of each crop attributable to pollination supply areas linked to it.
5. Aggregate and compare results from different scenarios and build a curve linking percentage of additional biodiversity protection to yield differential.
6. Use economic figures and/or purchasing power data to translate (5) into economic figures for decision makers.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

- Foundational ontologies, ensuring a common base for all others;
- Domain ontologies, describing the core biological, agricultural, ecological and physical phenomena;
- Authorities and vocabularies to describe the identities of crops, scientific taxonomies, geographical entities

B. How are they used? What for?

Within the platform, to programmatically access third party datasets and assist the assemblage and validation of a computational workflow that produces the answer to the query.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Ontologies are developed for the project and shared among the servers involved in producing the workflow They can be re-published online, as OWL resources. Authorities and vocabularies are reused resources from independent projects (e.g. GBIF, AGROVOC…), accessible online through the correspondent URLs and web services.

D. Do they come with a clear license? Which license?

Open license (CC BY 4.0 for ontology content, GNU Affero General Public License 3.0 for software). Data resources licensing terms vary.

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

ARIES (http://aries.integratedmodelling.org/) based on k.LAB (http://www.integratedmodelling.org)

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

For the process sketched above (see problem statement), the following critical points and needs emerge:

Evolving terminology in ontologies must be kept aligned with the systems consuming them, and users should be kept aware of the evolution. The terminologies in ontologies may evolve and become obsolete, the need arises then to replace them. The replacements can be accessed in an automated way, however, systems that consume the ontology should be aware of how this is

handled so they can notify users. Term status (pending curation, final, etc) should be exposed to the users

When linking foundational and domain ontologies, the need for parsimony in ontologies often conflicts with the speed with which results can be achieved: providing rich terminologies may ease locating a needed concept while making the overall knowledge landscape much more confusing and eventually impossible to navigate. **A mechanism to guide the user through the semantic layers and ensure usability without redundancy is needed.**

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

The ARIES platform only requires that the datasets to access are online. Most datasets used are spatial coverages in either raster or vector formats. Compatibility of most common data formats is enforced by the AI-supported engine.

Continuing the example given above, the following types of datasets would be required:

- Cropfields, including their spatial extent and the crop type;
- Areas with high biodiversity value or protected;
- Country-dependent yield data by crop and income or percentages of GDP attributable to each.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Decision-makers of various extraction, interested in exploring socio-economic system problem areas (these users interact with the semantic resources in ways that hide the semantics itself). Scenario developers (who need knowledge of the semantic underpinnings to properly annotate their inputs). Modellers of domain ontologies and supporting data and model components.

Organization and role

What is your organization and your role in it?

Basque Centre for Climate Change (BC3), Bilbao, Spain

Ikerbasque Research Professor

UC08: Lack of support for managing/finding/validating/reconciling/accessing alignments between ontologies by Clement Jonquet (LIRMM, University of Montpellier)

Problem statement

One of the challenges when dealing with multiple ontologies is the overlap and mappings between these ontologies.

Often alignments between ontologies are not given by the ontologies' maintainers, so when needed, they are created by users, sometimes with duplications but always with burden. When this happens, it may be unclear how to locate and choose among different sets of alignments between the same ontologies, and it is also difficult to reconcile existing sets of alignments. In general, the support for creation/validation/publishing/maintenance of alignments between ontologies is way less extensive than support for the same operations with ontologies.

We would like to address this subject by developing ontology mapping capabilities to align AgroPortal ontologies and make AgroPortal the reference platform for mapping extraction, generation, validation, evaluation, storage and retrieval by adopting a complete semantic web and open linked data approach and engaging the community.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

Mostly ontologies, logically defined (with axioms), with definitions, multilingual, mappings.

B. How are they used? What for?

Various uses.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

The ontologies used are all publicly accessible, with URLs.

D. Do they come with a clear license? Which license?

Open license or no license.

Semantic toolkit

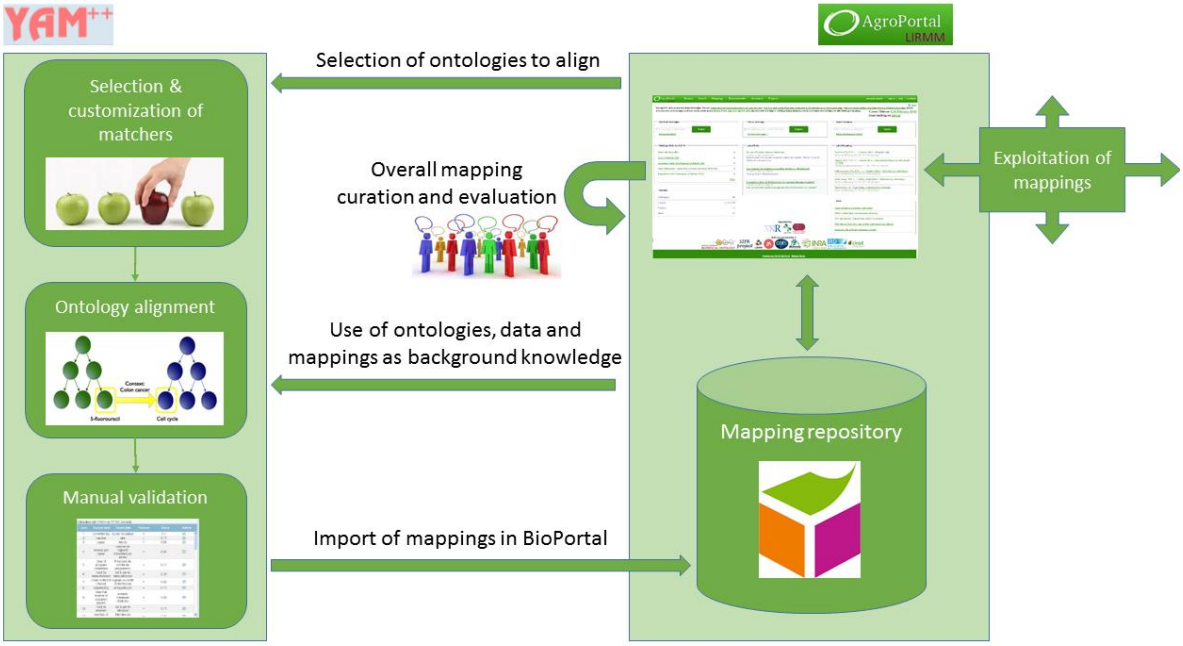
What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

AgroPortal (platform for hosting & serving agronomical ontologies): http://agroportal.lirmm.fr/

Yam++ (mapping tool): http://yamplusplus.lirmm.fr/

Currently, alignments are managed by different tools, assembled in pipelines. The goal is to have the needed functionalities integrated into a single workflow, together with other actions that may be performed on ontologies.

Our goal is to develop a complete ontology framework, based on AgroPortal and YAM++, that will capture the whole ontology mapping life cycle as illustrated in Figure.



Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Surprisingly, it seems there is a gap between the state-of-the-art results obtained at each edition of the Ontology Alignment Evaluation Initiative (OAEI - http://oaei.ontology matching.org) and the day-to-day reality of ontology developers. Tools are often hardly reusable, and results cannot be easily reproduced outside of the benchmarking effort.

Another key role of ontology repositories, such as AgroPortal, is to store mappings between ontologies. Ontology repositories shall support the extraction, generation, validation, evaluation, storage and retrieval of mappings between the ontologies they host. Automatic mapping generation within ontology repositories shall go beyond simple lexical or ID-based approaches and state-of-the-art tools shall be incorporated within repositories. An equivalent effort, such as the one made to harvest ontologies, must be made to harvest the mappings between these ontologies and describe them with metadata and provenance information to facilitate trust and reuse.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

N/a

We are seeking for ontology alignments in whatever format.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Ontology developers managing alignments. Users needing alignments. We also believe we should be able to engage the community in the mappings validation and evaluation steps. Maybe crowdsourcing could be an approach also.

Organization and role

What is your organization and your role in it?

Assistant Professor, LIRMM, University of Montpellier, France

UC09: *AgMIP Data Interoperability* by Cheryl Porter (AgMIP)

Problem statement

The Agricultural Model Intercomparison and Improvement Project (AgMIP) is a global community of agricultural modelers who are working to improve the ability of models to characterize global food insecurity due to climate change and other stressors. Model intercomparison and assessment activities use ensembles of models, running the same data sets to test the models' responses to climate-related inputs and to apply the models in scenario-based analyses to provide insight on impact, vulnerability and adaptation. The use of multiple models poses challenges because model data requirements differ from model to model. In addition, sources of data used in models vary substantially in quality, quantity, format and accessibility. The AgMIP IT Team has developed data interoperability methods and tools to facilitate AgMIP multi-model activities. Researchers have applied these tools extensively in ensemble modeling activities and the tools have been adopted widely, even in research which is unrelated to AgMIP. Our goal is to use these tools and methods to aggregate datasets over many locations, years, and management practices to enable meta-analyses; model development, validation, and improvement; statistical analyses; and other types of quantitative analyses. Such capability would increase the value of existing and future datasets beyond the original research purpose.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

The backbone of the AgMIP Data Interoperability tools is the **International Consortium of Agricultural System Applications (ICASA) Data Dictionary**, which was developed as a means to comprehensively document field experiments. AgMIP data are irregular and vary considerably; from relatively low quality datasets (e.g., farm and household surveys) to high quality, high volume datasets (e.g., detailed field experiments from research institutions). For this reason, a **flexible, extensible format was required to standardize and store the data**. A JSON (JavaScript Object Notation) structure with some hierarchy and minimal relationships was created, with the variable names in the ICASA Data Dictionary providing the keys for each key-value pair in the object.

The ICASA Data Dictionary is a hierarchical dictionary. It has been mapped to the CropOntology and AgrO ontologies.

B. How are they used? What for?

Data which are keyed to ICASA variable names are used in model-specific translators which format the data to the very specific requirements for each model. For example, the DSSAT Cropping System Model uses ASCII text files as input to the model in specific formats. The DSSAT translator can take data from AgMIP files, annotated with ICASA variables, and create the files needed to run the DSSAT model. Similarly, translators have been developed for APSIM, EPIC, SarraH and several other models.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

All AgMIP data, tools, and software are publicly available.

Data: https://data.agmip.org/cropsitedb
Tools: http://tools.agmip.org/
Software: https://github.com/agmip
Additional Resources: http://research.agmip.org/display/it/Data+Interoperability

D. Do they come with a clear license? Which license?

AgMIP software is licensed under the BSD-3-Clause license.
The AgMIP database adopts the Open Data Commons Attribution License .
The AgMIP datasets are licensed under the Creative Commons Attribution 3.0 License .

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

AgMIP tools include **libraries of translators** that have been implemented in desktop applications, high performance computing applications, and workflow systems. Applications include translators to format data from external sources to AgMIP Harmonized format, t**ranslators to format data into model-specific formats, user interfaces for AgMIP ensemble modeling activities, graphical analysis of model outputs**, etc.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

On the data supply side, there are **limitations on the datasets which are available** for harmonization using the AgMIP methods and tools. Initially, we would like to have regular workshops to bring together data collectors and people with expertise on the use of AgMIP tools for harmonization for the purpose of collecting datasets which are relevant to quantitative analyses, including modeling.

We also need to **develop additional model-specific translators** which can be used to provide AgMIP harmonized data for use by models. There are currently 13 modeling groups, which have developed AgMIP translators for their models. Maintenance of these translators is required as models evolve and provide new capabilities.

AgMIP multi-model activities have focused on crop production models, such as DSSAT and APSIM. More work is needed to **describe the types of data needed for socio-economic models, pest and disease models, livestock and grazing models**.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Types of data: Weather, soils, plant genetics, farm management practices, climate scenarios, adaptive management.

Formats: AgMIP harmonized data uses JSON structures. Individual models require very specific formats including ASCII, XML, databases, Excel spreadsheets, etc.

Storage & size: From very minimal storage for site-based analyses to many terabytes for global analyses.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Research scientists.

Organization and role

What is your organization and your role in it?

Agricultural Model Intercomparison and Improvement Project (AgMIP)

UC10: Farm Data Storage and Access, and Field Data Observation by Catherine Roussey (Irstea)

Problem statement

We would like to share all data used for crop management in our experimental farms. This could be:

- data provided by sensor in the field.
- Weather data provided by weather station
- Pest Attack observation extracted from alert bulletin
- Plot description with their cultivar and geometry
- GPS path of tractor
- etc...

These data can be used for several research purposes like define Decisions Support System dedicated to farmer's activities: pest management control or irrigation or tasks planification. The goal is to identify which part of data should be made available freely and open to ease agriculture researches and improve agricultural data interoperability and which data should be kept private to preserve farmer security and privacy.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

I need data schema with good documentation in english at least (french will be appreciated). We reuse Semantic Sensor Network ontology (SSN), *Sensor-Observation-Sampling-Actuator ontology* SOSA, Dublin Core, FOAF,....

I also need normalised vocabulary or reference dataset to provide value for property (value vocabulary). For example a dataset that define all the units (QUDT) or weather phenomenon (SWEET), value vocabulary that define crop group (winter wheat) or cultivar (missing).

When I need to compute aggregation I need a SKOS with correct hierarchical or semantic relation.

All the defined elements should be associated to a dereferenceable URI.

B. How are they used? What for?

All these semantic resources will be used to publish on the LOD our data related to our farm. We also want to compute some aggregated indicator about our farm management (economic, environmental and so one). To compute this, aggregated indicators should use different types of information sources. Thus, we want to solve data interoperability issues: that is to say integrate data from multiple sources: sensors, farmer information system, national database, etc.

See for example our weather dataset that publish the observation of our weather station http://ontology.irstea.fr/pmwiki.php/Site/WeatherData

And the archive of agricultural alert bulletin http://ontology.irstea.fr/pmwiki.php/Site/BSV

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Due to the difficulty to find appropriate semantic resource sometimes we are forced to create our own semantic resources. We created a new French Crop Usage Thesaurus to describe french crop group relations. Wheat belongs to Cereale and so one… This thesaurus was used to annotate french agricultural alert bulletins. It is available as a LOD dataset http://ontology.irstea.fr/cropusage/page/FrenchCropUsage as well as on AgroPortal

D. Do they come with a clear license? Which license?

As far as possible we create them with a public and open license

For instance, French Crop Usage is under CC-BY license

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

We use Protégé for data schema definition and extension.

We use CORESE (https://corese.inria.fr/) to query locally our dataset and check the consistency.
We have created our own transformation module based on pattern transformation. Tools like datalift do not fit our user needs. For example datalift does not save the transformation process as a tasks workflow so we have to redefine all the transformation when one source is updated.

We published our dataset thanks to Jena (https://jena.apache.org/) and Virtuoso sparql endpoint.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Several standards already exist about agriculture: ICASA, AgMIP, Foodies etc.. some are mature enough but are not easily accessible on the web (ICASA, AgMIP). Some web link may be broken. Several versions exist on the web which one is the reference version ? Some are published on the web but need to be improved to be easily reused (FOODIES, AGRORDF). Lack of documentation to understand how to use the property. Usually a property has only a name but it is not enough to understand the meaning and the usage. AGRORDF comments when they exist are in german.

When we use a well documented schema published on the LOD sometimes it is not easy to understand the design pattern and best practices uses (SSN, SOSA).

Unclear versions, governance uncertainty (ex: QUDT), translations or derivation do not evolve with source

Need ways to interact with resource producers to get explanation or suggest improvement

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

The data may be presented in CSV or any computer format and should be translated in RDF. Although some transformation tools like datalift exist, we create our own workflow to triplify our data using transformation patterns. Because we have to rerun the transformation several times with different files having the same structure.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

2 Computer scientists with knowledge in semantic web technologies (RDF, RDFS, SPARQL)

For selecting the best data model the computer scientist should also have knowledge in first order logics.

Organization and role

What is your organization and your role in it?

Researcher at Irstea french national research institut for agriculture and environment

UC11: *Soil Data Interoperability* by Giovanni L'Abate (CREA-AA)

Problem statement

Integrating available soil (open) data to provide farmers with smart tools suggesting decisions on agricultural systems (DSS or Decision support systems), based on key soil properties and other environmental data.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

Classifications, vocabularies, Thesauri, KOSs

B. How are they used? What for?

To classify soil (samples).

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Public (http://artemide.art.uniroma2.it/vocbench2/#Concepts) and shared: http://202.45.139.84:10035/catalogs/fao/repositories/agINFRA#node/<http://soilmaps.entecra.it/kos>; http://soilmaps.entecra.it/en/soilcontrolledvocabulary.html

D. Do they come with a clear license? Which license?

Italian Open Data License(IODL v2.0) and Creative Commons License version 3.0 (CC-BY)

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

vocbench2; AllegroGraph WebView 4.11

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Due to limited funding, we would could not afford dedicated staff and resources to tasks such as lifting the classifications we use into web-oriented formats, loading them in software for maintenance and providing the services needed to integrate them into the desired applications. We started testing with a known editor for SKOS vocabularies (VocBench 2.4) but we faced the institutional and financial problems of setting up a server where the tool could be installed and tested. It would be good to have a cloud-based web-service that would allow us to maintain and expose the generated vocabulaires according to best practices and principles of open, linked, fair data (in human readable format and exploitable).

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Point/Polygon/Raster Data Datasets; Databases, WFS; WMS; WCS; OpenLayers; KML, No storage issues jet, Database management and tool deployment

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Domain expert; Database administrator; software developer

Organization and role

What is your organization and your role in it?

Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA) - Centro di Ricerca Agricoltura Ambiente (CREA-AA),

Technologist

UC12: Make your soil research data available, accessible, discoverable and usable by Giovanni L'Abate (CREA-AA)

Problem statement

European soil scientists are required to follow the INSPIRE directive when collecting and publishing their soil samples. INSPIRE prescribe a model for the information to collect, but it does not provide indication of how this information should actually be to stored electronically. This leads to different researchers using different data schemes which make hard the comparison and validation of data. The problem also shows when scientists want to use that data to support the research papers submitted to journals and published. Research data forms the backbone of research articles and provides the foundation on which scientific, technical and knowledge in general is built. Data journals publish datasets and their documentation ("data papers"). Although AGRIS discovers about 595,643 papers citing the word “soil” and 346,950 related to the AGROVOC “soil” concept, soil data, datasets or collections, (hereafter defined as “soil data”) are seldom published online in Data journals or repositories. Searching “soil data” on 14 portals, from 0 (in 4 cases) to 1812 results were found being the INSPIRE GEOPORTAL and Harvard Dataverse (910 results) the most responsive. Beside those are increasing numbers, just a really little part of such data could be considered really available, accessible, discoverable and usable. **Deploying a structured soil database semantically based could fill a lack and offer a tool useful both to researchers, Data journals, repositories, but also to technical experts and to knowledge in general fulfilling the scope of the next RDA meeting “From Data to Knowledge”, Berlin.**

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

B. How are they used? What for?

To develop intuitive, customizability, easy to use, interoperable tools.

D. Do they come with a clear license? Which license?

Italian Open Data License (IODL v2.0) and Creative Commons License version 3.0 (CC-BY).

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

among others.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Organization and role

What is your organization and your role in it?

Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA) - Centro di Ricerca Agricoltura Am-biente (CREA-AA),

Technologist.

UC13: Agricultural Science and Technology thesaurus by Xuefu Zhang (CAAS)

Problem statement

We have huge amount of agricultural information resources. However, it is hard to find the needed information effectively and timely for researchers and scientists in the area. In order to effectively organize and reveal the literature resources for the national agricultural users, and enhance agricultural information service ability, we created the super thesaurus. It is a combination of subject and category structure.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

The Super Science and technology thesaurus is a knowledge organization system for English literature and Chinese users, which has a three layer structure. The bottom layer is the base word library, which formats the vocabulary of the source word bank, and then gets the unified base word library.

The category layer in the Super Science and technology thesaurus are closely connected with the concept layer and the ontology layer. Through the category, the concept category may be clearly defined, and the concept is categorized. It include hierarchical relationships, synonymy, textual definitions, bilingual languages, to the related concepts, etc.

B. How are they used? What for?

The thesaurus main functions are:

1) The macro structure of the thesaurus can be controlled by the category system;

2) It provides a way to retrieve the literature from a subject or a professional point of view to make up for the deficiency of the subject heading;

3) The category system will focus the Thesaurus on subject, and it will benefit to the comprehensive concept selection and relation discovery;

4) The thesaurus provides a way to find the controlled vocabulary;

5) Through the thesauri, to realize the sequence of knowledge base.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

The thesaurus is shared resources within national level information institutes in China.

D. Do they come with a clear license? Which license?

Without license.

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

It is a standalone system.

4. Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

1) The relationship between words is not perfect and needs a lot of manual intervention to get the overall revelation of the relationship between concepts.

2) The lack of standards in the category system.

3) The mapping of category system and literature classification system is insufficient. It is necessary to formulate and improve the suitable category system according to the user's use habits.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

There is no special data type or format required.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Chinese librarians, information specialists, managers, reference staff, information technicians, and users are main interacting persons with this sematic resources.

Organization and role

What is your organization and your role in it?

Director, Ph.D & Prof.

Agricultural Development and Research Department

Agricultural Information Institute,

Chinese Academy of Agricultural Sciences

UC14: *Farmer and farming data for sustainability* by Amanda Moura (Solidaridad Network)

Problem statement

Our organization develops sustainability intervention projects with farmers around the world, in partnership with supply chain players in over 10 different commodities. Gathering and analyzing data is fundamental for developing the content and scope of our interventions, as well as assessing their efficacy and efficiency.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

We are currently developing two local taxonomies. The first is a taxonomy that classifies farming practices considered relevant for sustainability assessments. It includes mostly hierarchical/specific relationships and is trilingual (English, Spanish and Portuguese), with some textual definitions in a few cases. The second is another taxonomy, which structures the farmers, farms and farming systems' profile characteristics and hard data we collect, as well data format requirements. We are studying developing this second taxonomy into an ontology in the near future.

We also use an external resource, FAO's Agrovoc.

B. How are they used? What for?

Our local taxonomies are used to index the data we gather on field projects and support cross analysis and project evaluation. We use Agrovoc to index our digital library of support material on sustainable farming.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

At the moment, our taxonomies are local resources.

D. Do they come with a clear license? Which license?

N/A

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

We are using VocBench 3 with GraphDB to develop and maintain the taxonomies, and are currently developing a CMS that will consume their data for indexing.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Most difficulties we have had so far are related with the learning curve of using free solutions for developing our semantic resources - the demand for new technical skills can be time-consuming, and paid solutions are currently out of our budget.

Educating collaborators on semantics and semantic resources is our next challenge this coming year, and in this we feel the lack of access to case studies and shared experiences.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

We currently have quantitative and qualitative data on farming practices and farming profile of about 6.000 producers, collected through online and paper self-assessment questionnaires. This year we will also be collecting data using an online mobile application, as well as technical farming assessments from field extensionists. Our library has around 500 bibliographic resources, structured with Dublin Core. Indexing of both farmer and bibliographic data is done manually, and they are in different MySQL databases hosted on Amazon Web Services.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

For the development of our local resources, we have a taxonomist, subject matter specialists, programme managers and a data analyst.

In the use/application of these resources, we have a taxonomist, programme managers and an information analyst involved. Our back end developers interact with the resources for integration purposes.

Agrovoc is used by an information analyst and a knowledge manager.

Organization and role

What is your organization and your role in it?

Solidaridad Network is an international civil society organization dedicated to facilitating the development of socially responsible, ecologically sound, and profitable supply chains. My role is as a taxonomist in the Knowledge and Information Management team in the São Paulo, Brazil office.

UC15: Publication of Inspire-based agricultural Linked Data by Raul Palma (PSNC)

Problem statement

Publication of Linked Data in agriculture compliant with the INSPIRE specifications. Linked data is increasingly becoming a popular method for publishing data on the Web because of the benefits it can bring, such as Improved data accessibility by both humans and machines, e.g., for finding, reuse and integration, for discovering more useful data through the links, and for exploiting data with semantic queries.

The beneficiaries of this work would be, first of all farmers and farm associations, but it could also provide benefits to the public authorities.

The expected impact would be improved access to heterogeneous datasets used and produced by farmers through more intuitive interfaces, supporting them in farm management tasks and decision making. See a presentation about this in the Linked Open Data in Agriculture 2017 Workshop.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

We use FOODIE ontology. The ontology enables the representation of data compliant with the agriculture FOODIE data model in semantic format and their interlinking with established vocabularies and ontologies (e.g., AGROVOC). The goal of of this model was to define the application vocabulary covering the different categories of information dealt by the farm mgmt. tools/apps, particularly those used within FOODIE project but generic enough to be reused and extended for covering many scenarios. Additionally the model was aimed to be designed in line with existing standards and best practices. Accordingly the model was designed by reusing data specifications from the INSPIRE directive, which in turn are based on ISO/OGC standards for geospatial services and formats, thus applying the ISO/OGC-approach of modeling physical things, so-called ”features”. The INSPIRE specifications are defined as UML models and are available in different XML-based formats (e.g., GML, XMI) and as Enterprise Architect (EA) projects. As result, FOODIE data model was specified in UML by extending and specializing INSPIRE data model for Agricultural and Aquaculture Facilities (AF) [1]. The final model was consulted with experts from various institutions, such as EU DG JRC, EU Global Navigation Satellite Systems Agency (GSA), Czech Ministry of Agriculture, Global Earth Observation System of Systems (GEOSS), German Kuratorium für Technik und Bauwesen in der Landwirtschaft (KTBL), and received very positive feedback.

In order to build the ontology, it was necessary to transform or lift FOODIE data model into semantic format. This process was conducted semi-automatically by reusing existing tools, and adhering to the mapping rules for transforming geographic information UML models to OWL ontologies defined by the ISO 19150-2 standard. In particular, we used the ShapeChange tool and addressed several issues and customizations before and after the execution of the ShapeChange processor as described in [2].

Thus, in line with the data model, different agricultural-related concepts can be described and represented with the ontology, including agricultural facilities, crop and soil data, treatments, interventions, agriculture machinery, etc. The ontology can be used for different semantic tasks, such as semantic data integration, data semantization (transformation of (semi-)structured data to semantic format); ontology-based data access (e.g., accessing relational databases as virtual, read-only RDF graphs); publication of linked data compliant with INSPIRE specification, including the discovery of links with relevant datasets in the Linked Open Data cloud.

In addition to use concepts from INSPIRE and ISO standards, the ontology also reuses concepts from well-known ontologies and vocabularies like dublin core, geosparql, and prov-o. In the latest version (4.6.2), it defines 135 classes, 124 object properties and 77 data properties, with a total of over 2400 axioms. Each term has a label in english and a pointer (rdfs:isDefinedBy) to the source vocabulary (for imported terms). Additionally, the ontology defines property ranges and cardinality constraints.

B. How are they used? What for?

For this use case, the ontology is used for publication of agriculture linked data compliant with INSPIRE directive.

Data from FOODIE pilot was transformed into semantic format according to the ontology and published as linked data.

In particular data from two pilots were used during the use case:

- Precision viticulture (Spain)
  Delivered a web-based solution providing advisory services in different aspects related to winegrowing, like disease prevention, production estimation or harvesting schedule
- Open Data for Strategic and Tactical planning (Czech Republic)
  Delivered two main applications, one for farm telemetry and other for estimation of yield potential

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

The ontology is available in https://github.com/FOODIE-cloud/ontology and also from http://agroportal.lirmm.fr/ontologies/FOODIE

D. Do they come with a clear license? Which license?

dct:license <http://creativecommons.org/licenses/by/4.0/>

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

We used several tools/applications:

- D2RQ for transforming Relational Databases as Virtual RDF Graphs
- RDF for the representation of data
- Farming (FOODIE) ontology providing the underlying vocabulary and relations
- Virtuoso for storing the semantic datasets
- Silk for discovery of (some) links
- Sparql for querying semantic data
- Hslayers NG for visualisation of data
- Metaphactory for visualisation of data

The generated linked datasets can be accessed via:

Sparql endpoint: https://www.foodie-cloud.org/sparql

Faceted search endpoint: https://www.foodie-cloud.org/fct

Demo Visualization on Map: http://ng.hslayers.org/examples/foodie-zones/

Metaphactory instance: https://foodie.grapphs.com/resource/Start, demo visualisation: https://foodie.grapphs.com/resource/TerrasGauda-plot1

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Main difficulty was on defining the data model. The transformation into ontology was a little time consuming. Then, transformation of relational data into RDF was also time consuming to define

appropriate mappings, as this mapping definition file (also RDF) was done manually. However the tool itself performs very good, both for having virtual graph access or generating an RDF dump. There are other tools (besides D2RQ) that may provide simpler interfaces for mapping definitions.

Silk is having many limitations and is difficult to use with large datasets, especially from sparql endpoint. So, link discovery could be improved. There are other tools like LIMES that can be tested, although LIMES also uses somehow Silk.

Virtuoso performance is currently very good, after performing some tunning tasks, and includes very good support for geo-spatial functions. Note that the triplestore including this agriculture data, also includes many other linked datasets, some of them quite large like the Open Land Use dataset (OLU), Open Transport Map dataset (OTM), and Smart Points of Interest dataset (SPOI).

Dataset Name	***Graph in FOODIE endpoint***	***Source***	***Triples***
OLU**	http://w3id.org/foodie/olu#	Transformed from PostgreSQL	127,925,971
SPOI	http://www.sdi4apps.eu/poi.rdf	Source provided by WRLS, modified and fixed before loading	381,393,555
NUTS	http://nuts.geovocab.org/	Open Source	316,238
OTM***	http://w3id.org/foodie/otm#	Transformed from PostgreSQL	154,340,611

Dataset Name	***Graph in FOODIE endpoint***	***Source***	***Triples***
Hilucs classification	http://w3id.org/foodie/hilucs#	Transformed from PostgreSQL	397
Urban Atlas*	http://w3id.org/foodie/atlas#	Transformed from PostgreSQL	19,606,025
Corine*	http://w3id.org/foodie/corine#	Transformed from PostgreSQL	16,777,533
Eurovoc	http://foodie-cloud.org/eurovoc	Open Source	425,667
Emergel	http://foodie-cloud.org/emergel	CTIC	256,239

* Selected subsets (The ontologies generated are available from https://github.com/FOODIE-cloud/ontology)

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

The data requirement is to be compliant with FOODIE farming data model. It can be in relational format (database) or directly as RDF. The transformation from relational to RDF can be easily re-applied (the mapping definition file is available). Publication in virtuoso is also a simple task. Finding links with Silk can be complicated, if large datasets are used (as mentioned in the limitations). There is currently no limitations in terms of size or storage.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

The person interacting with the ontologies, and other semantic tools has to have good knowledge of semantic technologies. The person interacting with the final dataset published as linked data could be any final user (e.g., farmers in this case)

Organization and role

What is your organization and your role in it?

Poznan Supercomputing and Networking Center (PSNC), Poznan, Poland

Semantic technologies coordinator.

References

[1] INSPIRE Thematic WG Agricultural and Aquaculture Facilities. D2.8.III.9 data specification on agricultural and aquaculture facilities. http://goo.gl/eWi6rq, Dec 2013.

[2] Raúl Palma, Tomas Reznik, Miguel Esbrí, Karel Charvat, and Cezary Mazurek. 2015. An INSPIRE-Based Vocabulary for the Publication of Agricultural Linked Data. In Revised Selected Papers of the 12th International Experiences and Directions Workshop on Ontology Engineering - Volume 9557, Valentina Tamma, Mauro Dragoni, Rafael Gonçalves, and Agnieszka Ławrynowicz (Eds.), Vol. 9557. Springer-Verlag New York, Inc., New York, NY, USA, 124-133. DOI=http://dx.doi.org/10.1007/978-3-319-33245-1_13

UC16: Food safety model repositories by Matthias Filter (BfR)

Problem statement

Food safety as a global challenge requires efficient knowledge transfer between academia, business operators and governmental agencies. In Europe, a rich variety of useful models, software tools and databases for food safety risk assessment exists, but exchange of these kinds of information between different resources is currently extremely difficult and time consuming. Integration of models and modelling tools is vital to cope with the numerous existing and emerging food safety risk and challenges. Several European institutions specialized in food safety modelling and risk assessment (ANSES, BfR, DTU, EFSA) currently collaborate to establish a Risk Assessment Modelling and Knowledge Integration Platform (hereinafter referred to as RAKIP) where the term “knowledge” specifically refers to data and models relevant for risk assessment tasks. The development of a RAKIP portal would improve transparency in data- or model-based knowledge and facilitating the exchange of this knowledge between different software tools that are already available in each of the three institutions.

The foundation of efficient knowledge exchange is however the encoding of knowledge in a harmonized data format (called Food Safety Knowledge Markup Language - FSK-ML). A critical component of this FSK-ML is a harmonized, open, community-driven ontology on metadata describing all relevant information for future use of data and models. The establishment of such semantic resources would be of of utmost importance not only for research and risk assessment institutes in Europe and the whole world, but also be an important resource for software developers and third party service providers that want to support the work of food business operators, risk assessors and decision makers in the food safety area.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

Details on the semantic resources used currently can be found here:

https://foodrisklabs.bfr.bund.de/rakip-harmonization-resources/

Among others currently the following third party resources are used: SSD-CODE, FOODON, MIME, PMM-Lab, OpenFSMR, Bibliographic Ontology Specification, vCard

B. How are they used? What for?

These resources are used mainly as basis for controlled vocabularies. In some cases these resources define the metadata itself, as e.g. vCard, RIS-format

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

See link above

D. Do they come with a clear license? Which license?

Don't know

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

Google sheet….

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

- Technical resources for EASY !! joint management and maintenance of semantic resources are missing
- Currently it is very hard to integrate useful semantic resources as there are multiple different resources that provide e.g. ontologies in different formats (owl, rdf) that are not always compatible
- The reusability of semantic resources by software tools is not straightforward. A special challenge is the generation of GUIs in software tools, that allow the domain experts to annotate their knowledge using those ontologies or controlled vocabularies that are available.
- Standards and best practices to represent and exchange semantic resources are at least not known to us (we are no experts in ontology work)

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

This use case focus on the new data format “FSK-ML” which allows to store software script based mathematical models and data together with relevant metadata in a coherent file format. The files contains therefore software code (e.g. R scripts or Matlab code) that can be executed by proper software tools in order to make model-based predictions. ….

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Software developers, Risk assessors, Research scientists

Organization and role

What is your organization and your role in it?

Federal Institute for Risk Assessment, Germany. I’m a senior research scientist.

UC17: *High-throughput phenotyping* by Alice Boizet (Inra)

Problem statement

Food security requires to select better adapted plant species and varieties to global changes. A way to do that is to use high-throughput phenotyping.

There are 3 types of high-throughput phenotyping platforms : field, green house or omics platforms. They work with different species and different protocols. We need to link semantic concepts, especially between field and greenhouse to be able to compare data. We need semantic interoperability.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

We use these ontologies : Agrovoc/GACS, Plant Ontology, Crop Ontology, Unit Ontology, PATO, Trait Ontology, Phenome Ontology

B. How are they used? What for?

They are used in PHIS (Phenotyping Hybrid Information System) to link variables to semantic concepts and to link objects to platforms in order to standardize concepts used in phenotyping

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

They all are publicly accessible on http://agroportal.lirmm.fr/

D. Do they come with a clear license? Which license?

I don’t know

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

The tools used are :

- Protégé, RDF4j, corese to conceive ontologies

- yam++ (matching)

- SKOS (semantic relation ontology)

They are not integrated in a workflow, but work is underway.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

It's difficult to work with too large ontologies which lacks precision. For example, it is hard to find a precise link between plant and farm plot in Agrovoc.

Ontologies contain mostly hierarchical links but we also need horizontal links which are very useful for alignment. We need application ontologies to organize & reason on data. But these application ontologies need to be mapped to references ontologies to be understood by other people.

Semantic resources repositories (e.g. AgroPortal) are still addressing ontology specialists and are difficult to use by domain researchers (biologists…). Interfaces are too complex, terms used are jargon.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Environmental data (temperature for example)

Phenomic data (leaf area for example) which are measured from plant images

crop

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Geneticists, ecophysiologists, bioinformaticians, agronomists, biologists, statisticians...

Organization and role

What is your organization and your role in it?

Research Engineer in INRA (France)

UC18: Food Traceability with respect to foodborne pathogen outbreak investigations by Damion Dooley (University of British Columbia)

Problem statement

Improving Food Traceability will lead to positive public health outcomes and economic savings due to the precision and speed of tracing pathogen contamination back to its source. Public and animal health will benefit, and the trend to low-cost technology for testing and distribution tracking will reduce the costs to the agriculture industry and to society for resolving outbreaks, and will minimize or eliminate unnecessary food quarantine and wastage.

In the future, proactive testing at source (farm/field/abattoir) for pathogens may occur, but until then investigations are required to determine the pattern of outbreak. An investigation triggers interviews with patients to determine common food consumption and possible transmission points; at some point distribution of particular goods is then traced. Endemic but "slow-burn" outbreaks may require genomic typing in order for patterns to be apparent.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

We've developed a few ontologies to try to foster a controlled vocabulary that would be applicable throughout the farm-to-fork distribution network. Both our FoodOn food ontology (http://foodon.org/) and the GenEpiO genomic epidemiology ontology (http://genepio.org) are built on hierarchical facets or branches, including textual definitions, some synonymy and axiomatization, and especially for FoodOn, several relations that enable food composition to be described simply and adequately. We rely as much as possible on a community of ontologies found at OBOFoundry.org to supply anatomy, chemistry, geography, phenotypic quality, the relation ontology, units and measures, etc. vocabularies that pertain to food.

B. How are they used? What for?

Foodon is a year old, and so still emerging from experimental mode - it needs more axiomatization of component foods in order to be useful to a reasoner in outbreak analysis to deduce pathogen related food. GenEpiO, also quite recent, is just being introduced as a vocabulary with a food component for describing foodborne pathogen samples from farms, animals, and other environmental or anthropocentric (built) environments like restaurants. It is used in a proposed sequence repository metadata specification.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

http://foodon.org/
http://genepio.org
http://genepio.org/geem/form.html#GENEPIO:0002083 (may take a little while to load).

D. Do they come with a clear license? Which license?

Creative Commons CC-BY

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

We use protege to edit the resources; and a variation on Chris Mungall's Ontology Starter Kit Makefile to compile release versions; we use Ontofox to fetch updated ontology contents. Our own scripts convert ontology contents - and standards expressed in ontology - to a tool we're developing called GEEM , a shopping mart of ontology-driven components, viewable at http://genepio.org/geem/ .

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Although OBOFoundry has solved some orthogonality problems between ontology domains that ideally describe the world cohesively, many problems remain. The OWL 2.0 formal logic, open world platform requires great training to wield appropriately with respect to reasoning, the ultimate test of data harmonization. The training barrier is preventing adoption of ontology-driven technology by many agencies/projects. Can we provide ontology terminology to foster global standardization without the training burden that currently entails that implementers of a vocabulary pretty much have to have the skills of a maintainer of the same vocabulary if they are trying to do more than just fetch picklists of terms.

There is too much variance at more abstract levels of relationship structure work to encourage global harmonization at that level. We are focusing on the lower-level fruit of establishing hierarchies of terms according to physical and categorical dimensions - such vocabularies can be adopted without disagreement engendered by more abstract (upper level ontology driven) designs. As well, some resources just aren't supported enough, like the Gazetteer ontology.

(I would be interested in a kind of "term spider" that could reach across databases and ontologies, establishing the lateral equivalency of terms like "Canada" or "gastrointestinal disease"; imagining with enough connections that it could support data mining; so far I've seen most work on the problem of mapping one ontology to another.)

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

OWL ontology, 10 to 30mb files (Gazetteer, Foodon, are relatively big, even if culled to fit a particular agency's context). Editing workflow is informal at moment; not many partners involved in each ontology upkeep.

Manpower

What is the profile of persons interacting with the semantic resource or their application?

On staff, directly, 2, indirectly 2; our recently started lab has a lead ontology developer, supported by 3 postdoc/doctoral student positions and a project coordinator and liaison to other agencies. As well our ontologies are being reused via ontology search engines although we don't have stats on that usage - would be good to encourage folks like Ontobee and OLS to enable stats on term lookups for

particular ontologies.) Larger reuse of foodon is happening via https://github.com/enpadasi/Ontology-for-Nutritional-Studies

Organization and role

What is your organization and your role in it?

University of British Columbia, Department of Pathology

UC19: *Professional Society member needs* by Stella Dextre Clarke (ISKO)

Problem statement

I shall try to respond from the point of view of a professional society whose members are interested in semantic resources and want to learn more. Although we have members worldwide, I shall be thinking particularly if those in the UK.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

Our members differ in the extent to which they already use these things. Some are comfortable with ontology development and manipulation, including all the features you mention, while others handle only simple taxonomies or authority lists, with little more elaborate than synonym control and hierarchical relationships…

B. How are they used? What for?

The traditional use of controlled vocabularies for indexing is still widely practised with collections of images and other non-text resources, but seems to be gradually diminishing elsewhere. Other uses are emerging, e.g. as Linked Data hubs across user communities or applying high-level ontologies to enabling discovery functions across a large enterprise…..

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Some of our members (like CABI) manage publicly available vocabularies, e.g. HASSET at https://hasset.ukdataservice.ac.uk/ , while others keep their vocabularies strictly in-house……

D. Do they come with a clear license? Which license?

Practice varies greatly, but open access without charge is becoming more common than before.

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

Our more experienced members use software packages such as Multites, Smartlogic Semaphore or Poolparty. Some have developed their own. But many others struggle on with Excel and Word….

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

For most of our members the biggest difficulty seems to be persuading management that money spent on vocabulary development, maintenance and use is a worthwhile investment. To address this we need to do more promotion of the success stories, showing outcomes that carry weight in the Boardroom….

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

In our meetings we try to cater for operations of all sizes….

Manpower

What is the profile of persons interacting with the semantic resource or their application?

Generally, not high enough! In the information professions we seem to attract quiet people, who enjoy doing good work that serves users well, but who lack personal ambition in climbing up the power ladder. Our people come from a wide range of backgrounds, including any field in the sciences or humanities, who somehow find themselves managing information access, information assets and information flows. Few of us are computer scientists and few would describe themselves as ”techies”….

Organization and role

What is your organization and your role in it?

Vice-President, ISKO; Vice-Chair, ISKO UK….

UC20: *AquaDiva* by Alsayed Algergawy (FSU Jena)

Problem statement

In the AquaDiva project, we are going to develop a provide a semantic layer on the top of existing data sets in order to facilitate and enhance data discovery and integration. To this end, we make use of semantic technologies in different places in the framework. For example, to cover the domain of the AquaDiva project, which is a multidisciplinary and covers a diverse number of domains, such biology, ecology, and others we develop a new ontology exploiting a number of existing ontologies on BioPortal.

Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

We make use of a set of existing ontologies (or parts) from bioportal

B. How are they used? What for?

We use them in order to develop our new ontology (AquaDiva ontology). First, we download the set of available ontologies on Bioportal. Then, we go to modularize (partition) each ontology, to select only relevant modules to our domain exploiting a set of domain terms collected from the project scientists.

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

Not yet, but we are going of course, to make it public

D. Do they come with a clear license? Which license?

Not yet

Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

Actually, we use different semantic tools and approaches, such as OWL API, bioportal service, and we also develop our own framework to put all together.

Limits and expectations

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

Of course, during building this framework we face a lot of challenges and problems starting from data upload to the repository till data discovery and integration. For example, during data upload, we need to annotate data attributes to be mapped to a concept in the Aquadiva ontology, however, till now this has been done manually which is a time consuming process.

Tell us a bit more about…

Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

Aquadiva supports dealing with data from different formats, e.g. tabular data (structured) as well as unstructured data (such as publications). The size of these data varies depending on the project providing this data set.